View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by KITopen

Compensation Approaches for Far-field Speaker Identification

Qin Jin, Kshitiz Kumar, Tanja Schultz, and Richard Stern

Carnegie Mellon University, USA

{qjin,kshitizk,tanja,rms}@cs.cmu.edu

Abstract

While speaker identification performance has improved dramatically over the past years, the presence of interfering noise and the variety of channel conditions pose a major obstacle. Particularly the mismatch between training and test condition leads to severe performance degradations. In this paper we explored several approaches to compensation for the effects of reverberation including compensation using linear post-filtering and frame-base score competition.

1. Introduction

Over the years automatic Speaker IDentification (SID) has developed into a rather mature technology that is crucial to a large variety of spoken language applications. However, SID systems still lack robustness, i.e. their performance degrades dramatically when the acoustic training data mismatch with the given test conditions [1][2]. Robustness is currently the major challenge for real-world applications of speaker recognition.

We proposed a reverberation compensation approach using cepstral post-filtering [8]. The goal of this processing was to minimize the squared distortion between training and testing features by passing the cepstral features that emerge from the feature extraction process through a linear filter. As noted above, the filter was designed to minimize mean square distortion between cepstral in the training and testing conditions. We also proposed the "Frame based Score Competition" (FSC) approach in [5] to improve speaker recognition in far-field situations. In this paper we further elaborate this approach by adding simulated data on a large variety of noise conditions, i.e. we artificially create additional data by applying a filter approach and extend the number and variety of models for the competition approach. The paper is organized as follows: In the next section we describe the reverberation compensation approach using cepstral post-filtering and experimental results on the YOHO database. Section 3 describes the Frame-based Score Competition approach and shows the experimental results on a live FarSID database and section 4 concludes the paper.

2. Reverberation Compensation using Cepstral Post-Filtering

2.1. Mathematical Representation of Reverberation

This section provides a representation of reverberated speech in terms of the corresponding clean speech. The SID system works in conventional fashion, by extracting features from the signal and determining which of a set of trained models provides the best match to an ensemble of incoming features. In order to maintain high SID accuracy, it is desirable that the features derived from reverberated speech closely match features derived from the corresponding clean speech. Let *x[n]* represent the cepstral features of a speech waveform (and not the original waveform itself), and let $y_u[n]$ represent the corresponding cepstral features after undergoing room reverberation. Because reverberation can be thought of as the convolution of the input speech with the effective impulse response of a room, there would be a constant difference between *x[n]* and $y_u[n]$ if these features represented long-term cepstra of the entire waveform. When *x[n]* and $y_u[n]$ are cepstral coefficients of brief segments of speech (as in short-time Fourier analysis), there is an interaction between the speech and the analysis window and the difference between *x[n]* and $y_u[n]$ is no longer constant. For simplicity, we propose that the reverberated cepstral features $y_u[n]$ can be represented as the convolution between the input cepstra *x[n]* and the cepstral coefficients *h[n]* representing the effects of the room:

$$y_u[n] = x[n] + \sum_{i=1}^{N_h} h[i]x[n-i] \quad (1)$$

Referring to Figure 1, $x = x[n]_{n=1}^{M}$ represents clean speech features and $y_u = y_u[n]_{n=1}^{M}$ represents the corresponding reverberated features. The subscript u in y_u indicates uncompensated speech in the testing environment. Thus, the assumption in (1) implies a linear filter in the cepstral feature domain with filter taps being $h = [1 h_1 \cdots h_{N_h}]$. Note that in this representation *h[0] = 1*. As a result of reverberation, system training will be performed on the features of clean speech *x* but testing will use reverberated features y_u. We define the instantaneous uncompensated distortion du[n] to be

$$d_u[n] = x[n] - y_u[n] = x[n] - h[n] * x[n] = \sum_{i=1}^{N_h} h[i]x[n-i] \quad (2)$$

(The second equality is valid because *h[0] = 1*.)

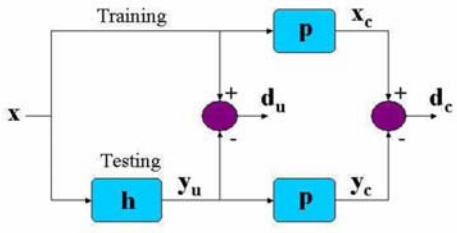

Figure 1: *Block diagram representing the model of reverberation and compensation*

We compensate for the effects of reverberation by imposing a finite-impulse response LTI filter on the observed features (*p* in Figure 1). We refer to the outputs of these features as *compensated*, and we use the notations x_c and y_c to indicate

features representing compensated clean speech and reverberated speech, respectively. We define the instantaneous *compensated distortion* $d_c[n]$ to be the difference between $x_c[n]$ and $y_c[n]$:

$$\begin{aligned} d_c[n] &= x_c[n] - y_c[n] = p[n] * x[n] - p[n] * h[n] * x[n] \\ &= p[n] * d_u[n] = \sum_{j=0}^{N_p-1} p[j] d_u[n-j] \end{aligned} \tag{3}$$

Where *Np* is the number of taps in the *p* filter. We seek to obtain the optimal *p* filter which, when applied to both *x* and y_u, minimizes the mean square compensated distortion $d_c[n]$ as defined above.

2.2. Solution to the Minimization Problem

In this section, we determine the optimal *p* filter as defined in Sec. 2.1. We define the objective for optimization to be the minimum expected distortion between the compensated training and testing features, and we find *p* to minimize $E\left[d_c^2[n]\right]$. Using (3), obtain $E\left[d_c^2[n]\right]$ as below:

$$\begin{aligned} \overline{d_c^2} &= E\left[d_c^2[n]\right] \\ &= \sum_{0 \le i,j \le N_p-1} p[i] p[j] E\left[d_u[i] d_u[j]\right] \end{aligned} \tag{4}$$

For evaluating $\overline{d_c^2}$ in (4), the terms $E\left[d_u[m] d_u[n]\right]$ can be obtained by using (2) as below:

$$E\left[d_u[m] d_u[n]\right] = \sum_{1 \le i,j \le N_h} E\left[h[i] h[j] x[m-i] x[n-j]\right] \tag{5}$$

Further, assuming that

$$\begin{aligned} &E\left[h[i] h[j]\right] = \sigma^2 \delta[i-j], \quad \sigma^2 \neq 0 \\ &E\left[h[i] h[j] x[m] x[n]\right] = E\left[h[i] h[j]\right] E\left[x[m] x[n]\right], \quad \forall i,j,m,n \end{aligned} \tag{6}$$

with δ being Kronecker delta, we can obtain $E\left[d_u[m] d_u[n]\right]$ in (5) as

$$E\left[d_u[m] d_u[n]\right] = N_h \sigma^2 R_x[n-m] \tag{7}$$

where *Rx* is the autocorrelation sequence of *x*. Substituting (7) into (4), we obtain

$$\overline{d_c^2} = N_h \sigma^2 \sum_{0 \le i,j \le N_p-1} p[i] p[j] R_x[i-j] \tag{8}$$

We can differentiate (8) with respect to *p* to find the optimal *p* but this will result in the optimal *p* being 0: if all the elements in *p* are equal to 0, all features in *x* and y_u will be mapped to 0, and the mean square distortion $\overline{d_c^2}$ will always be zero as well. While this is clearly the optimal solution in the mathematical sense, it is not a useful solution. In order to avoid the degenerate solution *p = 0* we further constrain *p*:

$$\sum_{j=0}^{N_p-1} p[j] \neq 0 \tag{9}$$

The constraint in (9) means that the *p* filter must have non-zero DC gain. Next, the fact that the p filter will be applied to both training and testing implies that scaling features by the same factor in both training and testing will leave the SID accuracy unchanged. This implies that we lose no generality by using the more specific constraint on *p*

$$\sum_{j=0}^{N_p-1} p[j] = 0 \tag{10}$$

To minimize $\overline{d_c^2}$ in (8) under (10), we construct a Lagrangian optimization criterion as below:

$$\begin{aligned} \Lambda(\mathbf{p}, \lambda) = N_h \sigma^2 \sum_{0 \le i,j \le N_p-1} p[i] p[j] R_x[i-j] + \\ \lambda\left(\sum_{j=0}^{N_p-1} p[j] - 1\right) \end{aligned} \tag{11}$$

Differentiating (11) with respect to $[p, \lambda]$ and equating the differentials to zero, we can obtain the optimal *p* as below:

$$\begin{bmatrix} R_x[0] & R_x[1] & \cdots & R_x[N_p-2] & 1 \\ R_x[1] & R_x[0] & & R_x[N_p-3] & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R_x[N_p-2] & R_x[N_p-3] & \cdots & R_x[0] & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \times \begin{bmatrix} p[0] \\ p[1] \\ \cdots \\ p[N_p-1] \\ \lambda' \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ 1 \end{bmatrix} \tag{12}$$

where $\lambda' = R_x[0] - \frac{\lambda}{N_h \sigma^2}$. Note that the unknown in $N_h \sigma^2$ due to the reverberation filter *h* has been incorporated into λ'. For later reference and compactness, we write (12) equivalently as (13).

$$\Phi \begin{bmatrix} \mathbf{p}^T \\ \lambda' \end{bmatrix} = \mathbf{b} \tag{13}$$

We note that *Rx*, the autocorrelation sequence of the clean features underlying the reverberated features, is the only unknown required to find *p* in (13), and specifically that the optimal solution does not depend on the reverberation filter h. We have thus designed an optimal post-filter *p* which is invariant to the reverberation due to *h* and thus far depends only on *Rx*. Of course, an SID system operating in a reverberant environment can observe directly the reverberated features y_u, but the autocorrelation of the clean speech Rx is not directly observable. Nevertheless, we can approximate Rx as the autocorrelation sequence obtained from clean features extracted from training data:

$$R_x[m] \approx R_T[m], \qquad m = 0, \cdots N_p - 2 \tag{14}$$

where $R_T[m]$ is the autocorrelation sequence obtained from clean features in training data. Combining (14) and (11), we can solve for *p*, so the *p* is invariant to the underlying clean features in *x*. The Φ matrix in (13) is Hermitian but not Toeplitz, its invertibility guarantees a solution that is both optimal and unique for *p*. Φ was found to be invertible in SID evaluations so combining (1), (6) (9), and (14), we claim that we have developed an optimal and unique solution for *p* which is not only invariant to the reverberation due to h but also invariant to the underlying clean features in *x*. This invariance greatly simplifies our SID system. We can design *p* using training features alone and use *p* to generate compensated training features x_c, as in Figure 1. The processed features x_c are used for training speaker models. During testing we apply the same filter *p* to the observed testing features in y_u and generate compensated testing features y_c, again as in Figure 1. Because the same *p* is applied across all reverberation conditions, no

modification of the filter design needs to be done for any particular reverberant environment.

2.3. EXPERIMENTAL VERIFICATION

2.3.1. Experimental Procedures

We developed the simulated FarSID database for this study, which was obtained by degrading clean speech from the English portion of Verbmobil I (VMI) database. The utterances consisted of high-quality recordings of spontaneously spoken face-to-face dialogs between two speakers who are discussing appointment scheduling and travel arrangements [12]. The data were recorded at a sampling rate of 16 kHz, with 16-bit resolution, and PCM encoding. For our study we selected a subset of 100 male speakers from the VMI database based on the durations of their utterances. The speakers' contributions are segmented by turns, where each file contains one turn. The turns are annotated for the dialog sessions to accommodate within- and between-session testing. In a second step we distorted the clean speech data by convolving it with simulated impulse responses that represent rooms of different size and reverberation time, and with differing distances between the sound source and the simulated microphones. The choice of the far-field conditions and the degradation process itself are described in detail in the next two subsections.

The far-field conditions used to compile this database were based on the results of a series of pilot studies. In these studies, we passed utterances from the YOHO database [7] through simulated room impulse responses of various types. The results of the pilot study were used to determine the ranges of the key parameters of room dimensions, reverberation time, and distance between source and microphone that were both compact and meaningful in terms of the type of reverberation introduced. We ultimately selected three different room sizes, four different reverberation times, and five different distances between source and microphone, which were observed to have a significant impact on speaker identification performance under simulated distorted conditions. The resulting far-field conditions are listed in Table 1. In total, the FarSID database contains 45 far-field conditions plus the original clean speech condition. We name each of the far-field conditions in the format <room><reverberation><distance>. For example, S2R05D100 refers to a medium conference room with reverberation time 0.5 seconds and a source-microphone distance of 100 centimeters.

Table 1. Far-field conditions for the FarSID Database

Room Type	Small (S1)	Medium (S2)	Large (S3)
Room size in meters (length, width, height)	(5,4,3)	(10,8,4)	(20,16,6)
Reverberation time (sec)	[0.3,0.5,1]	[0.3,0.5,1,2]	[0.3,0.5,1,2]
Source-mic distance (meters)	[0.5,1,2]	[0.5,1,2,4]	[0.5,1,2,4,8]
Total far-field conditions	9	16	20

We distorted the clean speech by convolving it with the simulated Room Impulse Response (RIR) generated using [6]. While the RIR program incorporates the dependence of room impulse response on many different physical characteristics of the room and environment, we focused on three major attributes, the room size, reverberation time, and distance between the source and microphone, using the parameter values in Table 1. We used the standard definition of reverberation time as the time time required for the acoustic signal power in the room to decrease by 60 dB when a sound source is turned off.

2.3.2. Experimental Results

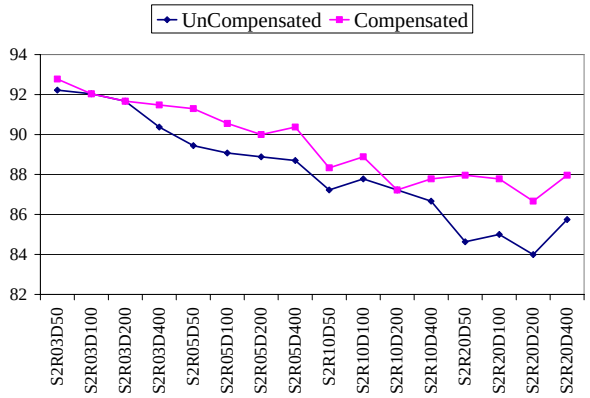

Figure 2: *SID performance in a medium-sized room with post-filtering reverberation compensation*

Figure 2 presents SID performance with the Reverberation Compensation algorithm applied across conditions in Medium Room. The compensation algorithm attempts to find an optimal causal linear filter which minimizes squared distortion due to mismatch in training and testing conditions. Under certain assumptions, the optimal filter is invariant to the RIR as well as invariant to the test speech under consideration. Compensation is applied individually to different cepstral features. The number of taps in the optima filter was taken as 5. The filter has to be applied during testing as well as in training and thus new speaker models need to be built by using filtered training speech. Comparing the results in Figure 5, we see that applying compensation algorithm provides substantial improvement in SID accuracy. We see that on an average across different conditions, applying compensation provides a relative reduction in error by 11% with the highest relative reduction being 21%.

2.4. DISCUSSION

In this section we discuss some of the assumptions made in this approach. At first we assumed a representation for reverberated features in (1). As features are generated by windowing on overlapping segments of speech signal, an exact relationship between reverberated and clean features is hard to find. We therefore, needed approximations and borrowed (1) from representation of reverberation as a LTI filer in time domain. $h_0 = 1$ was chosen to keep the problem analytically attractable. The assumption of (6) essentially means that the frequency response of the h filter is flat. This would occur if the RTs were constant over all frequencies. This assumption was made in simulating the speech data that was used for the results in Figure 2, but it is not empirically valid, as noted above. Nevertheless, the algorithm was also successful for the environmental conditions summarized in Figure 2. These data indirectly validate the proposition that the assumption of (14), although physically invalid, still can produce useful compensation results in practice. We can obtain an estimate of number of taps in the optimal p as the knee in the curve describing the dependence of $\overline{d_c^2}$ on Np.

Although $\overline{d_c^2}$ decreases with a larger number of taps, the dissimilarity among compensated features for different speakers also decreases. As *Np* becomes very large, the p filter converges to be a uniform moving average filter that smoothes out all the data and reduces every feature to its mean value, which is zero for the MFCC features under consideration. This reduces the SID decisions to a random guess. For these reasons, we expect a local maximum in performance as a function of *Np*. Next, we note that the optimization construction in section 2.2 guarantees that for optimal *p*, the mean squared distortion for compensated case $E\left[d_c^2[n]\right]$, is never greater than that for uncompensated case $E\left[d_u^2[n]\right]$. Further, the optimal *p* is a linear phase filter.

The post-filtering algorithm was also applied directly to the speech signal in the time domain but in this case uncompensated case outperformed compensated case. This indicates that the modeling and assumptions in Section 2.1 and 2.2 is not easily generalizable to the time domain.

Our approach for dereverberation is somewhat similar to Wiener Filtering at a conceptual level. The major digression from the Wiener filter is that the p is applied to both training and testing speech, which leads to different requirements and solutions to the problem. While our approach is invariant to the detailed nature of the reverberation this is not the case for Wiener Filtering. We will consider generalizations of our approach in later studies. We will also apply the algorithm for speaker verification tasks.

3. Frame-base Score Competition (FSC)

In this section we first quickly review the decision process of speaker identification systems based on GMM likelihood scores and then summarize our FSC approach which is described in more detail in [5]. Let S be the total number of enrolled speakers and $LL(X \mid \Theta_k)$ the log likelihood score that the test feature sequence X was generated by the GMM Θ_k of speaker *k*, which consists of *M* mixtures of Gaussian distributions. Then the recognized speaker identity S^* is given by:

$$S^* = \arg\max_k \{LL(X \mid \Theta_k)\} \qquad k = 1,2,\cdots,S \tag{15}$$

Since the vectors of the sequence $X = (x_1, x_2, \cdots, x_N)$ are assumed to be independent and identically distributed, the likelihood score for speaker *k* and model Θ_k is computed as: $LL(X \mid \Theta_k) = \sum_{n=1}^{N} LL(x_n \mid \Theta_k)$. Our multiple microphone setup allows us to build multiple GMM models $\Theta_k^{CH_i}$ for each speaker *k* and each channel CH_i, resulting in a set $\Theta_k = \{\Theta_k^{CH_1}, \cdots, \Theta_k^{CH_C}\}$ for *k* speakers and *C* channels. The key idea of our FSC approach is to use this set of multiple GMM models rather than a single GMM model for the speaker identity decision. In each frame we compare feature vector x_i provided by channel CH_h to the multiple GMMs of speaker *k*. The highest log likelihood score is chosen to be the frame score. In this case the likelihood score of the observed features given speaker *k* is computed as:

$$\begin{aligned} LL(X \mid \Theta_k) &= \sum_{n=1}^{N} LL(x_n \mid \Theta_k) \\ &= \sum_{n=1}^{N} \max_{j=1}^{C} \left\{ LL\left(x_n \mid \Theta_k^{CH_j}\right) \right\} \end{aligned} \tag{16}$$

Note that this process does not rely on models for the test channel. Also, this competition process differs from the mono-channel scoring process in that per-frame log likelihood scores for different speakers are not necessarily derived on the same channel.

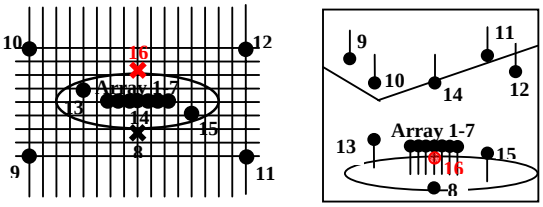

Figure 3: *Microphone setup in the FarSID database*

3.1. Data and Setup

A Far-Field Speaker Identification (FarSID) Database has recently been collected at Carnegie Mellon University to study the performance of speaker identification algorithms in adverse conditions, including a far-field microphone setup, various interfering noise sources, and reverberant room characteristics. Similar to the database described in [5] the FarSID database consists of speech recordings from multiple far-distant microphones as depicted in Figure 3. In addition, to make the FarSID database even more challenging than its predecessor database, we additionally recorded under various noise and reverberation conditions.

The FarSID database consists of conversational speech recorded in face-to-face dialog sessions under two different reverberation conditions (small vs. medium-sized room) under six noise conditions per room. The noise conditions were applied by playing interfering noises at different Signal to Noise Ratio (SNR) levels (music, white noise, speech) while recording the respective sessions. Each condition lasted for about 13 minutes per session per speaker. The different noise conditions and their respective SNR are indicated below:

- No noise
- Music Noise -5 dB SNR
- Music Noise 10dB SNR
- White Noise -5 dB SNR
- White Noise 10 dB SNR
- Speaker Interfering Noise -5 dB

In total we recorded 10 native speakers of American English, where each speaker is engaged by an interviewer in a conversation about various topics. Each speaker participated in 4 recording sessions, two in a small and two in a medium-sized room, totaling to 2 hours duration per speaker.

Figure 3 shows the distant microphone setup in the FarSID database. The right hand side illustrates the microphone positioning in the 3D space. Five microphones (labeled 9 to 12 and 14) are hanging from the ceiling or are mounted to high microphone stands. Seven microphones (labeled 1 to 7) are building a microphone array, with a distance of 5cm between each microphone. The microphone array and two other microphones (labeled 13 and 15) are set up on the table which is

arranged between the interviewer and interviewee. In addition, we use two lapel microphones, number 8 worn by the interviewer and number 16 worn by the interviewee, i.e. the speaker whose identity is to be recognized. The left hand side of Figure 3 illustrates the distance between the speaker (marked by a red "X") and these 16 microphones. One grid unit roughly corresponds to 0.25 meters.

3.2. Experimental Setup and Results

The experiments reported in this section are based on the FarSID database. The aim of the investigation is to demonstrate the robustness of our FSC approach for far-field scenarios and show how it addresses the challenges posed by mismatched condition, i.e. the fact that speaker models are applied to acoustic channel conditions which have not been observed during training. For this purpose we train and test our approach under four scenarios, where each scenario is designed to be more challenging than the previous one and more close to the challenges for real-world applications. The four scenarios are described next, followed by the experimental results in the respective subsections.

- **[Match]** Matched condition: train a speaker model with data recorded under the same conditions as in the test case – this is the golden line as it reflects the best case scenario.
- **[Mis-MM]** Mismatched condition with Multiple Microphone data: train speaker models with data simultaneously recorded by multiple microphones that cover a variety of conditions but the test condition.
- **[Mis-SM-k]** Mismatched conditions with Single Microphone data and knowledge about test condition: train speaker models with data recorded with one microphone under one condition and tested on a different condition using some prior knowledge about the test condition.
- **[Mis-SM-nk]** like [Mis-SM-k] but this time we varied the number of microphone positions involved for model training.

The experiments were carried out on data with the first noise condition of the FarSID database (see section 2), i.e. distant speech with common background noise such as air conditioning, computer fans, and reverberation recorded by the 16 microphones (as described above) in a medium-sized room. We selected 60 seconds of these data per speaker to train the speaker model. For testing we selected 30 seconds per speaker of the same noise condition and same room. The major mismatch results from the selection of the microphone positions for training and test, as described above. In total we had 106 test trials. All described experiments are conducted as closed-set speaker identification. Performance is measured in terms of identification accuracy, i.e. the percentage of correctly identified test trials. The applied speech features X are Mel Frequency Cepstral Coefficients (MFCC); the speaker models consist of Gaussian Mixture Models (GMM) with 64 Gaussians per model.

3.2.1. Experiment on [Match] Scenario

To get the upper bound performance, i.e. the best case scenario we trained and tested under the matched scenario. The second column of Table 2 shows the breakdown of speaker identification performance for each microphone position. To get the performance for microphone position y we trained all speaker models on the training trials recorded by microphone y and tested on the test trials recorded by microphone y. So, on average we get 98.4% identification rate on 10 speakers for far-field recordings if we assume that we know the test condition and that we do have recordings in this condition available for each speaker.

3.2.2. Experiment on [Mis-MM] Scenario

The results of the second experiment are described in column 3 of Table 2. Here we assume that we do have simultaneous recordings from microphone positions 9-15. To calculate the performance on position 9 we train 6 speaker models per speaker, one on each of the remaining positions 10, 11, 12, 13, 14, and 15. For the final number given in the table we average over the identification rates for each of these mismatching conditions. We achieve 92.1% accuracy over all positions, i.e. a drastic drop from the matched condition.

The fourth column in Table 2 compares this brute-force approach to our FSC technique. The same 6 models per speaker are now combined using competition at the scoring stage. As can been seen FSC significantly improves over the brute-force approach and even gets close to the [Match] performance. This result indicates that FSC compensates well for scenarios, in which recordings with multiple microphones but the matching one are available for a speaker. Our earlier results also showed that the SID performance further improves if the matching condition is available [5].

Table 2. *Performance for Multi-Microphone Setup*

Test Microphone	[Match]	[Mis-MM]	[Mis-MM] FSC
Position 9	98.1	85.8	97.2
Position 10	99.1	91.2	99.1
Position 11	98.1	92.1	97.2
Position 12	96.2	90.4	97.2
Position 13	100.0	94.2	100.0
Position 14	99.1	95.9	99.1
Position 15	98.1	95.0	98.1
Average	**98.4**	**92.1**	**98.2**

3.2.3. Experiment on [Mis-SM-k] Scenario

In this next set of experiments we investigate the performance of our FSC approach in the more realistic case where only single microphone recordings are available per speaker. We assume here without loss of generality to have training data from microphone position 4 [Mis-SM4]. As column 3 of Table 3 (labeled as [Mis-SM4]) shows, the performance drops drastically compared to the matched and the multi-microphone performance. The gap is more substantial for microphone positions 9 – 12, which is intuitively clear as these are further away from microphone 4 than microphone positions 13 – 15.

The key idea to make use of our FSC approach in the single microphone case is to simulate multiple microphone recordings from the single microphone data. In order to apply FSC, we simulated different channels from the microphone 4 speech. This simulation targets microphone positions 9-15 by convolving the source speech with the simulated Room Impulse Response (RIR) generated using [6]. While the RIR program incorporates the dependence of room impulse response on many different physical characteristics of the room and environment, we focused here on three major attributes, the room size,

reverberation time, and distance between the speaker and the microphone.

Column 4 in Table 3 (labeled [Mis-SM4]-FSC) shows the performance when simulating the 7 microphone positions. FSC on simulated channels significantly outperforms the baseline under mismatched conditions although it cannot beat the performance of FSC on real multi-microphone data "[Mis-MM] FSC" from Table 2. Please note that for both, the FSC on real multi-microphone data and on simulated multi-microphone data, we purposely exclude the data from the matching channel, i.e. we assume to not know the microphone position of the test condition. However, we do assume in the simulation to have some knowledge about possible microphone positions, i.e. the RIR filter do know the actual room size and reverberation time, and create realistic microphone distances.

Table 3. *Performance for Single-Microphone Setup*

Test Microphone	[Match]	[Mis-SM4]	[Mis-SM4] FSC
Position 9	98.1	57.5	84.9
Position 10	99.1	66.0	93.4
Position 11	98.1	68.9	93.4
Position 12	96.2	71.7	94.3
Position 13	100.0	94.3	97.2
Position 14	99.1	95.3	98.1
Position 15	98.1	93.4	97.2
Average	**98.4**	**78.2**	**94.1**

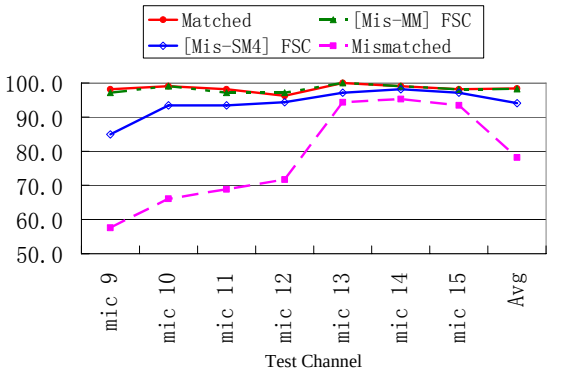

Figure 4: *SID Performance comparison under all setups*

Figure 4 summarizes our findings on the four cases, i.e. trained and tested on microphone 4 [Match], real multi-microphone conditions with FSC [Mis-MM]-FSC, single microphone conditions with simulation [Mis-SM4]-FSC, and the mismatched case [Mismatched], in which the models are trained on single microphone data at position 4 and applied to position 9-15 microphones.

3.2.4. Experiment on [Mis-SM-nk] Scenario

In this final set of experiments, we compared the impact of the number of microphone positions on performance. We tested this by repeating the [Mis-SM4]-FSC experiments but this time applying FSC on different numbers of simulated multi-microphone data streams. FSC6 refers to the case, in which we used all 6 mismatched microphone data (position 9 -15 except the position matched with test condition) to train 6 models. This corresponds to experiment [Mis-SM4]-FSC. FSC5 refers to the case where we used only 5 out of 6 simulated multi-microphone data. Since we can have 6 over 5 = 6 different choices, we averaged the performance over all different choices. In addition, we calculated the best and worst performance depending on the choice. FSC4, FSC3, FSC2 repeat the same experiments with fewer microphones giving us 6 over 4 = 15, 6 over 3 = 20, and 6 over 2 = 15 choices, respectively. Figures 5 and 6 show the best and worst performance for all selections.

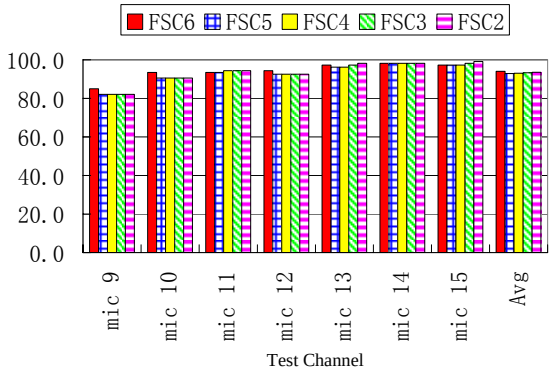

Figure 5: *Best performance of [Mis-SM4]-FSC*

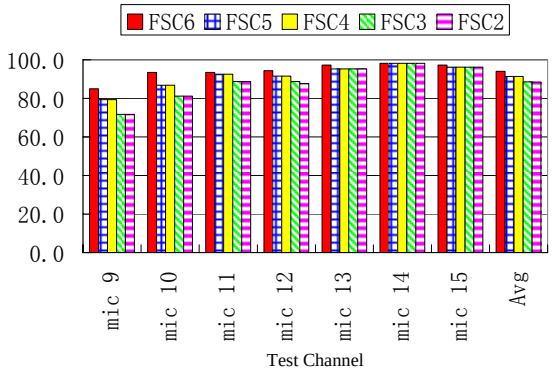

Figure 6: *Worst performance of [Mis-SM4]-FSC*

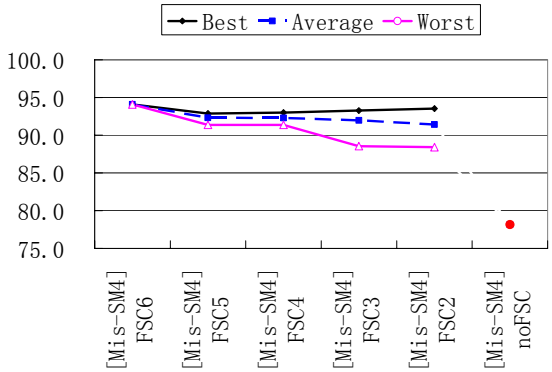

Figure 7: *Performance summary of FSC with different number of channels*

As can been seen from Figures 5 and 6 the worst selection of microphone positions for the data simulation does not have a significant impact on the system performance compared to the best choice. In other words, the success of the FSC approach does not depend on a proper selection of microphone positions for the simulation. Figure 7 compares the worst, average, and best case and with the mismatched condition. Even in the worst case, FSC still significantly improves performance compared to the baseline performance under mismatched condition (noFSC in Figure 7)

4. Discussion and Conclusions

In this paper, we presented an algorithm for reverberation compensation using Cepstral Post-Filtering. The compensation procedure consisted of a relatively simple FIR filter that is applied to sequences of cepstral coefficients, with the coefficients of the filter optimized to minimize the mean square difference between the compensated coefficients for speech in the training and testing environments. The optimal filter obtained was unique and invariant to the environmental conditions of a particular test trial. This approach provided significant improvement in SID accuracy across different reverberation conditions encompassing simulated as well as actual RIRs and also across different speech databases.

In this paper we also reported far-field speaker recognition performances under mismatched conditions. The aim of the investigation is to demonstrate the robustness of our frame-based score competition approach (FSC) for far-field scenarios and show how it addresses the challenges posed by mismatched condition, i.e. the fact that speaker models are usually applied to acoustic channel conditions which have not been observed during training. For this purpose we trained and tested our approach under four scenarios, where each scenario is designed to be more challenging than the previous one and more close to the challenges for real-world applications. The first scenario assumes to know the test condition, i.e. the best but most unlikely case. The second case assumes to not know the test condition but to have training samples recorded from multiple microphone positions for the speaker in question. Here, FSC significantly improves over the mismatched case, i.e. applying multi-microphone data gives more robustness since they cover multiple microphone positions and thus better prepare for the unknown. In the third scenario we assume to have only single-microphone data available and compensate this lack by simulating multi-microphone data using room impulse response filters. FSC manages to still significantly outperform the mismatched scenario. In other words, even when only single microphone recordings from a speaker are available, the simulation of multiple microphone recordings combined with our FSC approach improves the overall performance significantly. In the last scenario we vary the selection of microphone positions for the simulated data and show that even if we make the worst choice of microphone positions, we still see significant improvements over the mismatched case.

5. Acknowledgements

The work was funded in part by the National Geospatial-Intelligence Agency (NGA) under National Technology Alliance (NTA) Agreement Number NMA 401-02-9-2001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NGA, the United States Government, or Rosettex. We would also like to thank Fred Goodman for useful discussions and comments.

6. References

[1] Chin-Hui Lee, Frank K. Soong, Kuldip K. Paliwal, “Automatic Speech and Speaker Recognition: Advanced Topics”, Springer, 1996, ISBN:0792397061.

[2] S. Furui, “Towards Robust Speech Recognition Under Adverse Conditions”, ESCA Workshop on Speech Processing in Adverse Conditions, p. 31-42, 1992.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz and D. A. Reynolds, “A Tutorial on Text-Independent Speaker Verification,” Journal on Applied Signal Processing 4, pp. 430-451, 2004.

[4] D. Reynolds, “Speaker Identification and Verification Using Gaussian Mixture Speaker Models,” Speech Communication, Vol. 17, No. 1-2, p. 91-108, August 1995.

[5] Q. Jin, T. Schultz, and A. Waibel, “Far-field Speaker Recognition,” IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No.7, p. 2023-2032, 2007.

[6] D. Campbell, K. Palomäki, and G. Brown, “A MATLAB Simulation of “Shoebox” Room Acoustics for use in Research and Teaching. http://cis.paisley.ac.uk/research/journal /V9/V9N3/campbell.doc

[7] J.P. Campbell and D.A. Reynolds, “Corpora for the evaluation of speaker recognition systems,” IEEE ICASSP, 1999.

[8] K. Kumar, and R. M. Stern, “Environment-Invariant Compensation for Reverberation using Linear Post-Filtering for Minimum Distortion”, IEEE International Conference on Acoustics, Speech, and Signal Processing, April 2008, Las Vegas, Nevada.

[9] S. G. McGovern, “A model for room acoustics,” http://2pi.us/rir.html.

[10] CMU Sphinx Open Source Speech Recognition Engines, http://cmusphinx.sourceforge.net/html/cmusphinx.php.

[11] RWCP Sound Scene Database in Real Acoustical Environments, http://tosa.mri.co.jp/sounddb/indexe.htm.

[12] S. Burger, K. Weilhammer, F. Schiel, H.G. Tillmann, “Verbmobil Data Collection and Annotation, in: Wolfgang Wahlster (Ed.) Verbmobil: Foundations of Speech-to-Speech Translation, Springer-Verlag Berlin, Heidelberg, New York, pp. 537-549, 2000.